PA31A-1822

Findability of Federal Research Data

J.A. Hourclé
NASA-GSFC (Wyle)
joseph.a.hourcle@nasa.gov

Virtual Solar Observatory
http://virtualsolar.org/

Although many federal agencies have been providing access to scientific research data for years if not decades, the findability of the data has been quite lacking. Many discipline-wide efforts have been made in the big science communities, such as PDS for planetary science and the VOs in night-time astronomy and heliophysics, but there is a lack of single entry point for someone looking for data.

The `science.gov` website contains links to many of these big-science search systems, but doesn't differentiate between links to science-quality data and websites or browse products, making it more difficult to search specifically for data.

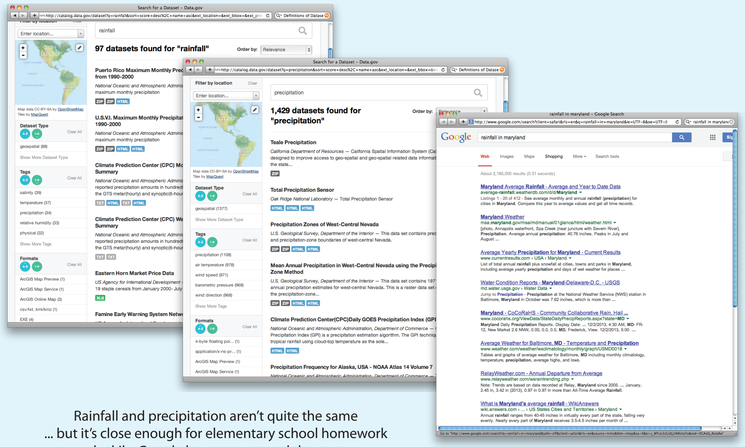
The `data.gov` website is a useful repository for PIs of small science data to stash their data, particularly as it allows interested parties to interact with tabular data. Unfortunately, as each group thinks of their data differently, much of what's now in the system is a mess; collections of data being tracked as individual records with no relationships between them. Big science projects also get tracked as single records, potentially with only a single record for missions with multiple instruments and significantly different data series.

We present recommendations on how to improve the findability of federal research data on `data.gov`, based on years of working on the Virtual Solar Observatory and within the science informatics community.

Similar / Equivalent Data

Users are given no indication if there might be similar concepts. Without knowing the proper jargon from the community that created the data, average citizens may not realize data may be available. Use of a taxonomy, thesaurus, ontology or similar for indexing would allow citizens to find data from related concepts or drill down into narrower concepts. Rolling up records into collections would reduce the amount of cataloging that would need to be performed.

Even without the overhead of cataloging, a vocabulary system can be used to expand queries using synonyms and equivalent terms.



Rainfall and precipitation aren't quite the same
... but it's close enough for elementary school homework
... luckily, Google knows to expand the query

Inconsistent Granularity

Even with 'dataset' not having a consistent meaning across communities, `data.gov` seems to also use it to mean 'website' or 'ftp directory', and neglects tracking the diverse, heterogeneous data that may be available from a given archive.

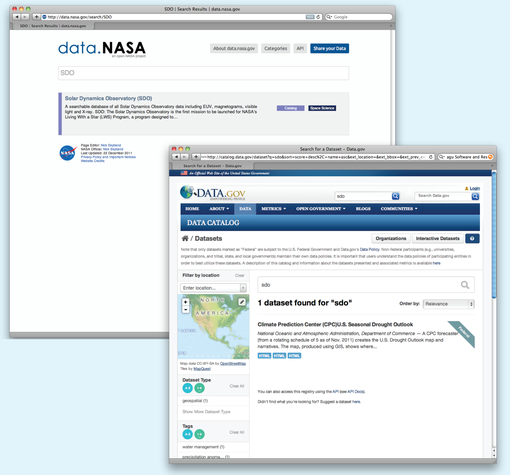
The link from `data.gov` sends you to a list of missions
that hasn't been updated in over 5 years
... but the SDAC tracks over 100 different types of data from
different instruments and processing applied, spanning decades
... including data that's only a few minutes old

Incomplete Inventory

Citizens have no way of knowing if data exists that may not yet be ingested.

In cases such as `data.nasa.gov`, no public call was put out to ask the community what data was available; one agency admitted that they used internet search engines to find websites with data. This resulted in major gaps and inconsistent granularities of data as a website might link to data from multiple instruments or even multiple missions.

In the case of agencies that conduct research, every named project, mission, experiment, investigation, grant or similar should have some record in the system, even if it's to give an estimate on when the data will be available. These 'placeholder' records could be inserted while data is going through official channels to be ingested and validated.



SDO is listed at `data.nasa.gov`, but ***not*** at `data.gov`
... and there are over a dozen sites that serve SDO data.

Excessive Returns

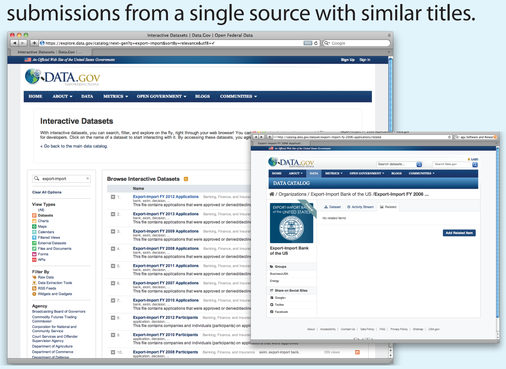
Many groups release related data as discrete records; they may be a record for each year, or records for each state or other region.

There needs to be an easy way to find and select:

- Complete time-series for a given measurement
- Data from other regions, especially those in adjacent regions.

Assigning records into collections of similar data or collected with the same methodology allows users to easily grab all data of interest without scrolling through pages of similar data.

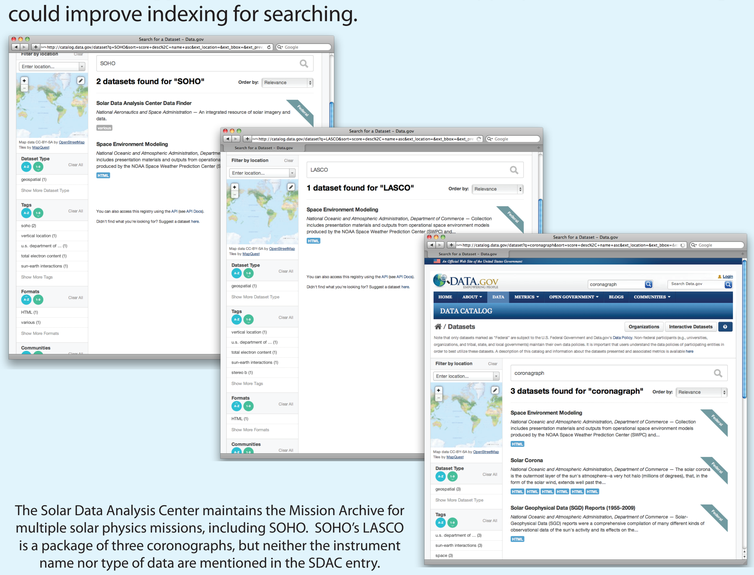
Much of this could be accomplished by looking at data submissions from a single source with similar titles.



The Import-Export Bank has time series of two similarly titled products
... but none have the others listed as being related.

Annotation / Cataloging

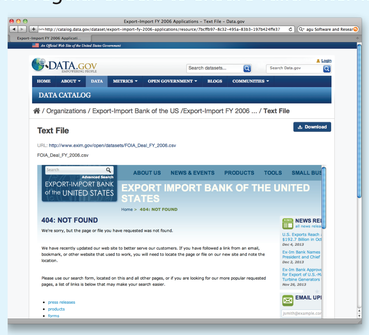
Data owners and managers know how both citizens and their designated communities refer to their data. Coordinating with them would ensure that common phrases used to search for their data are used within the system. Allowing the public to suggest annotation to records (with appropriate editorial oversight) could improve indexing for searching.



The Solar Data Analysis Center maintains the Mission Archive for multiple solar physics missions, including SOHO. SOHO's LASCO is a package of three coronographs, but neither the instrument name nor type of data are mentioned in the SDAC entry.

Maintenance

`data.gov` doesn't archive all of its data; it also includes a registry (`catalog.data.gov`). In cases where the data is not being managed locally, the system should perform regular checks to ensure valid external links.



attempting to view the file returns a 404 error

References:

Hourclé (2008a) "Data Relationships: Towards a Conceptual Model of Scientific Data Catalogs", AGU Fall Meeting, abstract #IN22A-03. `http://virtualsolar.org/frbr/`

Hourclé (2008b) "Reconciling Heterogeneous Data Catalogs", 2008 CODATA. `http://virtualsolar.org/frbr/`

Renear, Sacchi & Wickett (2010) "Definitions of Dataset in the Scientific and Technical Literature", *Proc. ASIS&T* `http://dx.doi.org/10.1002/meet.14504701240`